BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37563
doi: 10.3897/biss.3.37563

OPEN ACCESS

Conference Abstract

Infrastructure for Long-term Preservation and OCR Analysis of Herbarium Images

Nicolas Cazenave ‡

‡ CINES, Montpellier, France

Corresponding author: Nicolas Cazenave (cazenave@cines.fr)

Received: 20 Jun 2019 | Published: 02 Jul 2019

Citation: Cazenave N (2019) Infrastructure for Long-term Preservation and OCR Analysis of Herbarium Images. Biodiversity Information Science and Standards 3: e37563. https://doi.org/10.3897/biss.3.37563

Abstract

Herbaria hold large numbers of specimens: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. High resolution digital images of these specimens take up substantial bandwidth and disk space. New methods of extracting information from the specimen labels have been developed using OCR (Optical character recognition) techniques, but the exploitation of this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, alongside the variation and age of the fonts. Much of the information is handwritten and natural handwriting pattern recognition is a less mature technology than OCR. Today, our system (eTDR-European Trusted digital Repository) provides the OCR technology (using Tesseract software) adapted to the requirements of herbarium specimen images and requires minimal installation in each institution. This is what we propose to make available to botanists with our portal.

The goal for a museum is to be able to submit a large number of scanned images easily to a long-term archiving system in order to automatically obtain OCR texts and retrieve them by a full text search on an open data portal.

Most of the images are provided for reuse through CC-BY licenses. In each case, the rights of reuse associated with the data are specified in associated metadata.

This pilot was an opportunity to test the long-term storage service eTDR provided by CINES. The services (B2SAFE, B2Handle) developed by EUDAT were used to facilitate the transfer of data to the storage repository and to provide indexing services for access to that repository.

This workflow that has been tested for the european project ICEDIG is presented as a poster: See the document (Suppl. material 1).

Keywords

biodiversity data, herbarium, digitization, optical character recognition, data reuse, biodiversity infrastructure

Presenting author

Nicolas Cazenave

Hosting institution

CINES (National Computing Center for Higher Education) is a French public institution, located in Montpellier, France and supervised by the French Ministry for Higher Education and Research.

Supplementary material

Suppl. material 1: Infrastructure for long-term preservation and OCR analysis of herbarium images doi

Authors: Nicolas Cazenave
Data type: Poster
Download file (7.16 MB)